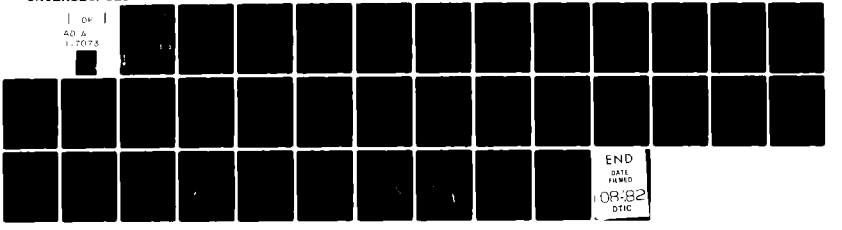
AD-A117 073 TEXAS A AND M UNIV COLLEGE STATION INST OF STATISTICS F/G 12/1
TIME SERIES MODEL IDENTIFICATION, SPECTRAL ESTIMATION, AND FUNC--ETC(U)
JUN 82 E PARZEN N00014-82-MP-2001
UNCLASSIFIED TR-N-31 NL



AD A117073

TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843-3143

UTE OF STATISTICS
713 - 845-3141

(12)

TIME SERIES MODEL IDENTIFICATION, SPECTRAL ESTIMATION, AND FUNCTIONAL INFERENCE

by Emanuel Parzen

Institute of Statistics

Texas A&M University

Technical Report No. N-31

June 1982

Texas A&M Research Foundation
Project No. 4226T

"Multiple Time Series Modeling and
Time Series Theoretic Statistical Methods"

Sponsored by the Office of Naval Research

DTIC
ELECTE
JUL 20 1982
S D
F

Professor Emanuel Parzen, Principal Investigator

Approved for public release; distribution unlimited.

FILE COPY

82 07 19 009

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER N-31	2. GOVT ACCESSION NO. AD-A117073	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) TIME SERIES MODEL IDENTIFICATION, SPECTRAL ESTIMATION, AND FUNCTIONAL INFERENCE		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Emanuel Parzen		8. CONTRACT OR GRANT NUMBER(s) ONR N00014-82-MP-2001 ARU DAAG 29-80-C-0070
9. PERFORMING ORGANIZATION NAME AND ADDRESS Texas A&M University Institute of Statistics College Station, TX 77843		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE June 1982 13. NUMBER OF PAGES
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Time series analysis, model identification, spectral estimation, ARMA model, information divergence, functional inference.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This survey talk seeks to emphasize the following ideas:

1. Functional inference formulation of parameter estimation
2. Parameter estimation and information theory
3. Information divergence of spectral density functions
4. Model identification, prediction theory, and memory
5. ARMA model identification for short memory time series
6. Model identification of long memory time series
7. The array of spectral estimators
8. Quantile approach to non-Gaussian time series analysis

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

TIME SERIES MODEL IDENTIFICATION, SPECTRAL ESTIMATION, AND FUNCTIONAL INFERENCE

Emanuel Parzen
Institute of Statistics
Texas A&M University
College Station, TX 77843 USA

Statistical spectral analysis has a literature which expands constantly. Every collection of papers about signal processing contains many papers discussing and extending the methodology of spectral analysis. Researchers in every scholarly discipline attempt to apply spectral analysis to problems in their field. Consequently I believe that the disciplines of time series analysis and spectral analysis need to be developed in a general way that is capable of treating problems in all the diverse fields in which they occur.

This survey talk seeks to emphasize the following ideas:

1. Functional inference formulation of parameter estimation;
2. Parameter estimation and information theory;
3. Information divergence of spectral density functions;
4. Model identification, prediction theory, and memory;
5. ARMA model identification for short memory time series;
6. Model identification of long memory time series;
7. The array of spectral estimators; and
8. Quantile approach to non-Gaussian time series analysis.

To be presented at Workshops on Signal Processing in the Ocean Environment, April/May, 1982. Research supported in part by the Office of Naval Research.

1. Functional Inference Formulation of Parameter Estimation

A general formulation of statistical theory as methods of analysis of statistical data assumes that a statistical question starts with a probability model for the observed data set, or sample, which is a function of a parameter to be estimated;

$$f(\text{sample}|\text{parameter})$$

represents the probability density function of the sample as a function of the parameter.

Classical statistical inference assumes that the parameter is a finite dimensional vector $\theta = (\theta_1, \ldots, \theta_k)$. Functional inference assumes that the parameter is a function, such as $f(\omega)$, $0 \leq \omega \leq 1$.

Parameter estimators are best determined by one of two general statistical principles: Bayes theorem or minimum divergence.

Bayes theorem assumes a prior distribution for the parameter, and computes the posterior distribution of the parameter given the sample, denoted

$$f(\text{parameter}|\text{sample}) \quad .$$

A minimum divergence method introduces a function $I(f;f_\theta)$ which measures the divergence between the true density f and a proposed density f_θ. If one can form a raw estimator $\tilde{f}$ of the probability density of the sample, then an estimator $\hat{\theta}$ is determined by minimizing $I(\tilde{f};f_\theta)$ with respect to θ. We call $\hat{\theta}$ a

minimum divergence estimator. When θ is a finite dimensional vector, $\hat{\theta}$ often exists; we call $f_{\hat{\theta}}$ a parametric-exact estimator of the true f. When θ is a function we have two main approaches to determine estimators $\hat{\theta}$ which we call non-parametric penalty and parametric-select.

A non-parametric penalty method determines $\hat{\theta}$ by minimizing

$$I(\tilde{f};f_{\theta}) + \lambda J(\theta)$$

where $J(\theta)$ is a measure of the smoothness of the function (tending to ∞ as θ becomes less smooth) and λ is a penalty parameter which is chosen by the researcher to balance the fidelity measure $I(\tilde{f};f_{\theta})$, assumed to tend to 0 as θ becomes less smooth, *with the smoothness measure* $J(\theta)$.

A parametric-select method determines $\hat{\theta}$ by approximating θ by a vector θ_m of suitable dimension m, called the order, and setting $\hat{\theta} = \hat{\theta}_m$ where $\hat{\theta}_m$ minimizes over θ_m the divergence $I(\tilde{f};f_{\theta_m})$. The problem of selecting the "best" order m is a problem of model identification; the problem of then estimating $\hat{\theta}_m$ can often be treated as a problem of classical statistical estimation of a finite dimensional vector.

When an estimator, denoted $\hat{\theta}$, is used as an estimator of θ, one has to take into account two kinds of errors, called respectively bias and variance. Bias is a measure of the deterministic difference between θ_m and θ, while variance is a measure of the stochastic distance between $\hat{\theta}_m$ and θ_m. As m increases bias decreases while variance increases. This is an

example of the fundamental problem of empirical spectral analysis which is how to achieve an optimal balance between bias and variance. [When one uses autoregressive spectral estimation, this problem reduces to a question of determining the order m of the approximating autoregressive scheme.]

2. Parameter Estimation and Information Theory

A general approach to determining divergence measures is provided by information theory. Let f(y) and g(y) be two probability densities on the real line, $-\infty<y<\infty$. The information divergence of index α of a (model) g from (a true density) f is defined for $\alpha = 1$ (index 1) by

$$I_1(f;g) = \int_{-\infty}^{\infty}\{-\log \frac{g(y)}{f(y)}\}\, f(y)\, dy$$

and for $\alpha>0$ (but $\alpha\neq 1$) by

$$I_\alpha(f;g) = \frac{-1}{1-\alpha} \log \int_{-\infty}^{\infty} \{\frac{g(x)}{f(x)}\}^{1-\alpha} f(x)\, dx$$

Information divergence of index 1 has a preferred role because it has an important decomposition

$$I_1(f;g) = H(f;g) - H(f)$$

defining

$$H(f;g) = \int_{-\infty}^{\infty}\{-\log g(y)\}\, f(y)\, dy,$$

$$H(f) = H(f;f) = \int_{-\infty}^{\infty}\{-\log f(y)\}\, f(y)\, dy.$$

We call $H(f;g)$ the cross-entropy of f and g, and $H(f)$ the entropy of f.

Information divergence of index 1 is usually referred to just as information divergence.

Another fundamental decomposition concerns the information $I(Y|X)$ about a continuous random variable Y in a continuous random variable X, defined by

$$I(Y|X) = I_1(f_{Y|X};\ f_Y)$$

$$= E_X I_1(f_{Y|X=x}; f_Y)$$

The entropy of Y and conditional entropy of Y given X are defined by

$$H(Y) = H(f_Y)$$

$$H(Y|X) = H(f_{Y|X}) = E_X H(f_{Y|X=x})$$

One can show that

$$I(Y|X) = H(Y) - H(Y|X).$$

Accession For
NTIS GRA&I
DTIC TAB
Unannounced
Justification
By
Distribution/
Availability Codes
Dist | Avail and/or Special
A
DTIC COPY INSPECTED 2

To apply these concepts to the problem of identifying a probability model for a sequence of random variables Y(1), Y(2), ..., we define the information divergence between the probability densities for the infinite sequence by

$$I_1(f;g) = \lim_{T\to\infty} \frac{1}{T} E_f \left[-\log \frac{g(Y(1),\ldots,Y(T))}{f(Y(1),\ldots,Y(T))}\right]$$

3. Information Divergence of Spectral Density Functions

We next consider a time series $Y(t)$, $t=0, \pm 1, \ldots$ which is a zero mean Gaussian stationary time series with covariance function

$$R(v) = E[Y(t)Y(t+v)] \quad ,$$

correlation function

$$\rho(v) = \frac{R(v)}{R(0)} = \text{Corr } [Y(t),Y(t+v)]$$

Despite the possible confusion with a probability density, we use f to denote the spectral density function

$$f(\omega) = \sum_{v=-\infty}^{\infty} e^{-2\pi i v\omega} \rho(v), \qquad 0\leq\omega\leq 1.$$

assuming $\sum_{v=-\infty}^{\infty} |\rho(v)| < \infty$. The frequency variable ω is usually assumed to vary in the interval $-0.5\leq\omega\leq 0.5$. But only the interval $0\leq\omega\leq 0.5$ has physical significance. We prefer the interval $0\leq\omega\leq 1$ for mathematical reasons.

A theorem of Pinsker (1963) can be interpreted as providing a formula for the information divergence between the probability density of a zero mean Gaussian stationary time series (normalized by its variance) with (true) spectral density $f(\omega)$, and the probability density of a zero mean Gaussian

stationary time series (normalized by its variance) with (proposed model) spectral density $f_\theta(\omega)$:

$$I_1(f(\omega);\ f_\theta(\omega)) = \frac{1}{2} \int_0^1 \{\frac{f(\omega)}{f_\theta(\omega)} - \log \frac{f(\omega)}{f_\theta(\omega)} - 1\} \quad d\omega$$

It has an information decomposition:

$$I_1(f(\omega);f_\theta(\omega)) = H(f(\omega);f_\theta(\omega)) - H(f(\omega)),$$

defining cross entropy

$$H(f(\omega);f_\theta(\omega)) = \frac{1}{2} \int_0^1 \{\log f_\theta(\omega) + \frac{f(\omega)}{f_\theta(\omega)}\} \quad d\omega$$

and entropy

$$H(f(\omega)) = \frac{1}{2} \int_0^1 \{\log f\ (\omega) + 1\}$$

From a time series sample Y(t), t=1,2,...,T one can form a raw estimator $\tilde{f}(\omega)$ of the true spectral density $f(\omega)$ by

$$\tilde{f}(\omega) = \left| \sum_{t=1}^{T} Y(t)\ e^{-2\pi i\omega t} \right|^2 \div \sum_{t=1}^{T} Y^2(t)$$

The sample correlation function

$$\hat{\rho}(v) = \sum_{t=1}^{T-v} Y(t)\ Y(t+v) \div \sum_{t=1}^{T} Y^2(t)$$

is computed by (for $0 \leq v < Q-T=M$)

$$\hat{\rho}(v) = \frac{1}{Q} \sum_{k=0}^{Q-1} \exp\ (2\pi i \frac{k}{Q} v)\ \tilde{f}(\frac{k}{Q}) \quad .$$

Estimating the parameters θ [of a parametric model $f_\theta(\omega)$] by minimizing the information divergence

$$I(\tilde{f}(\omega); f_\theta(\omega))$$

or equivalently the cross-entropy

$$H(\tilde{f}(\omega); f_\theta(\omega))$$

is asymptotically equivalent to the method of maximum likelihood.

An important example of the foregiong general considerations is the autoregressive model (of order m) for a spectral density. It has parameters $\theta = (\sigma_m^2, \alpha_m(1), \ldots, \alpha_m(m))$ and is defined by

$$f_m(\omega) = \sigma_m^2 |\ g_m(e^{2\pi i\omega})|^{-2} \quad ,$$

where

$$g_m(z) = 1 + \alpha_m(1)z + \ldots + \alpha_m(m)\ z^m$$

has all its roots in $|z| \geq 1$, the exterior of the unit circle in the complex z-domain. Then

$$2H(\tilde{f}; f_m) = \sigma_m^2 + \frac{1}{\sigma_m^2} \int_0^1 \tilde{f}(\omega) |g_m(e^{2\pi i\omega})|^2\ d\omega$$

is minimized by $\hat{\sigma}_m^2$, $\hat{\alpha}_m(1),\ldots,\hat{\alpha}_m(m)$ satisfying the sample Yule-Walker equations. The autoregressive spectral estimator

$$\hat{f}_m(\omega) = \hat{\sigma}_m^2 \, |\hat{g}_m(e^{2\pi i\omega})|^{-2}$$

is a parametric-exact estimator when the time series Y(t) obeys an autoregressive scheme of order m, and is a parametric-select estimator when the autoregressive scheme is adopted as an approximating model.

Maximum entropy characterization of AR spectra. The spectral density $f(\omega)$ that maximizes entropy $H(f(\omega))$ among all $f(\omega)$ satisfying the constraints

$$\int_0^1 e^{2\pi i\omega j} f(\omega)\, d\omega = \rho(j), \quad j=0,\pm 1,\ldots,\pm m$$

for specified correlations $\rho(j)$ is the autoregressive spectral density $f_m(\omega)$ with coefficients determined by the Yule-Walker equations. A "one-line" proof of this fundamental fact, originally stated by Burg, is as follows: from

$$H(f;f_m) = \frac{1}{2}\{\log \sigma_m^2 + 1\}$$

it follows that

$$0 \leq I_1(f;f_m) = H(f;f_m) - H(f)$$
$$= H(f_m) - H(f)$$

and

$$H(f) \leq H(f_m).$$

The foregoing simple proof of the maximum entropy character of autoregressive spectral densities is analogous to a proof of

the maximum entropy character of exponential model probability densities.

4. Model Identification, Prediction Theory, and Memory

Discussions of general statistical principles are usually concerned with the principles of parameter estimation. The more important problem of model identification does not yet receive the systematic attention and emphasis merited by its crucial importance. I believe we are in a position to describe qualitatively the types of models that are usually fitted to "Gaussian" time series that are analyzed by methods related to spectral analysis. We distinguish 4 model types which we call

1. No memory or white noise
2. Short memory or stationary
3. Long memory (or non-stationary)

3a. Long memory: transform to short memory

3b. Long memory: long memory plus short memory.

In the definition and identification of these models, we use the ideas of prediction theory. The information about a time series $Y(t)$ at time t in the m most recent values $Y(t-1),\ldots,Y(t-m)$ is denoted

$$I_m = I(Y|Y_{-1},\ldots,Y_{-m})$$

$$= I(Y(t)|Y(t-1),\ldots,Y(t-m))$$

For a Gaussian stationary time series

$$I_m = -\frac{1}{2}\log \sigma_m^2$$

where

$$Y^{\mu,m}(t) = E[Y(t)|Y(t-1),\ldots,Y(t-m)]$$

$$Y^{\nu,m}(t) = Y(t) - Y^{\mu,m}(t)$$

$$\sigma_m^2 = E[|Y^{\nu,m}(t)|^2] \div E[|Y(t)|^2]$$

As m tends to ∞, I_m tends to

$$I_\infty = I(Y|Y^-) = I(Y(t)|Y(t-1),\ldots)$$

$$= -\frac{1}{2}\log\sigma_\infty^2 = -\frac{1}{2}\int_0^1 \log f(\omega)\,d\omega = -H(f(\omega)) + \tfrac{1}{2}.$$

Further, if $H(f)>-\infty$, then $H(f_m)\to H(f)$ and $I(f;f_m)\to 0$.

We then define a time series Y(t) to be

no memory: $\sigma_\infty^2 = 1$, $I_\infty = 0$

short memory: $0 < \sigma_\infty^2 < 1$, $0 < I_\infty < \infty$

long memory: $\sigma_\infty^2 = 0$, $I_\infty = \infty$.

In terms of the dynamic range of the spectral density $f(\omega)$,

$$DR(f) = \max_{0\le\omega\le 1} f(\omega) \div \min_{0\le\omega\le 1} f(\omega)$$

we define intuitively

no memory: $DR(f) = 1$

short memory: $1 < DR(f) < \infty$

long memory: $DR(f) = \infty$.

The models we build for a time series depend on its memory type. A model corresponds to a transformation of the time series to a white noise series. Therefore a no memory or white noise time series requires no further modeling, although one may be interested in determining such statistical characteristics as the mean, variance, and probability distribution.

5. ARMA Model Identification for Short Memory Time Series

A short memory time series $Y(t)$ is modeled by an invertible filter which transforms it to white noise:

$$Y(t) \longrightarrow \boxed{\begin{array}{c}\text{innovations} \\ \text{filter } g_\infty\end{array}} \longrightarrow \varepsilon(t) = Y^\nu(t)$$

The infinite memory prediction errors are denoted

$$Y^\nu(t) = Y(t) - Y^\mu(t) = g_\infty(L)Y(t),$$

$$g_\infty(z) = 1+\alpha_\infty(1)\ z+\ldots+\alpha_\infty(m)z^m+\ldots,$$

$$L\ Y(t) = Y(t-1)$$

We call $Y^\nu(t)$ the innovation series and it is a white noise time series with variance $\sigma_\infty^2 R(0)$. In general a short memory time series is modelled by representing it, or approximating it, by an ARMA (p,q) scheme:

$$Y(t) + \alpha_p(1)\ Y(t-1)+\ldots+\alpha_p(p)\ Y(t-p)$$

$$= \varepsilon(t) + \beta_q(1)\ \varepsilon(t-1)+\ldots+\beta_q(q)\varepsilon(t-q)$$

where the polynomials

$$g_p(z) = 1+\alpha_p(1)z+\ldots+\alpha_p(p)\ z^p$$

$$h_q(z) = 1+\beta_q(1)z+\ldots+\beta_q(q)\ z^q$$

are chosen so that all their roots in the complex z-plane are in the region $\{z:|z|>1\}$ outside the unit circle. Then $g_p(z)$ and $h_q(z)$ are the transfer functions of invertible filters. $\varepsilon(t)$ is assumed to be a white noise time series which we identify with the innovations $\varepsilon(t) = Y^{\nu}(t)$;

$$\sigma^2_{p,q} = E[\varepsilon^2(t)] \div E[Y^2(t)]$$

is an estimator of σ^2_{∞}. The spectral density of an ARMA(p,q) scheme is

$$f_{p,q}(\omega) = \sigma^2_{p,q}\ \frac{|h_q(e^{2\pi i\omega})|^2}{|g_p(e^{2\pi i\omega})|^2}$$

The process of identifying ARMA (p,q) schemes which are adequate (and parsimonious) approximating models for a time series can be studied rigorously, and various at least semi-automatic methods are available which are based on order determining schemes.

The conditions under which the exact (or true) model is an AR(p) or ARMA(p,q) can be stated in terms of information

numbers. Define the information about Y in X_2 conditional on X_1 by

$$I(Y|X_1;X_1,X_2) = H(f_{Y|X_1}) - H(f_{Y|X_1,X_2})$$

$$= H(Y|X_1) - H(Y|X_1,X_2).$$

Then Y is AR(p) is equivalent to (where Y^- denotes the infinite past $Y_{-1}, Y_{-2},\ldots$)

$$0 = I(Y|Y_{-1},\ldots,Y_{-p};Y^-) = I_\infty - I_p \quad ;$$

Y is ARMA(p,q) is equivalent to

$$0= I(Y|Y_{-1},\ldots,Y_{-p}, Y^\nu_{-1},\ldots,Y^\nu_{-q}; Y^-)$$

Given a time series sample, Y(t), t=1,2,...T, of length T, one can calculate successively (using Fast Algorithms such as the Yule-Walker equations) estimators

$$\hat{I}_p = -\frac{1}{2}\log\hat{\sigma}_p^2, \quad p=1, 2, \ldots \quad .$$

which can be regarded as test statistics for testing white noise, or more precisely AR(0) against AR(p). The work of Akaike (1974, 1977) and Hannan and Quinn (1949) leads one to conjecture that a universal test for white noise (whose theory needs further study) is of the form (for a suitable choice of constant $c\geq 0$, say $c=1$)

$$- \frac{1}{2} \log \hat{\sigma}_p^2 - \frac{p}{T} \log \log T \leq \frac{c}{T} \text{ for } p=1, 2, \ldots \quad .$$

A related conjecture is that optimal orders p of approximating autoregressive schemes can be identified by first determining the orders at which are attained the absolute and relative minima of order determining criteria which determine orders p for which $\hat{I}_\infty - \hat{I}_p$ is not significantly different from zero.

Akaike's order determining criterion AIC is defined by

$$AIC\ (m) = \log \hat{\sigma}_m^2 + \frac{2m}{T} \quad ;$$

it seeks to determine the order of an exact autoregressive scheme which the time series is assumed to obey. One can raise the objection against it that it does not consistently estimate the order, which is done by a criterion due to Hannan and Quinn (1979):

$$AICHQ(m) = \log \hat{\sigma}_m^2 + \frac{m}{T} \log \log T$$

Parzen (1974), (1977) introduced an approximating autoregressive order criterion called CAT (criterion autoregressive transfer function), defined by

$$CAT(m) = \frac{1}{T} \sum_{j=1}^{m} (1-\frac{j}{T}) \hat{\sigma}_j^{-2} - (1- \frac{m}{T}) \hat{\sigma}_m^{-2} \quad .$$

One chooses the value of CAT(0), such as

$$CAT(0) = -(1 + \frac{1}{T}),$$

in order to accept the hypothesis of white noise when it is true a specified percentage (say 90%) of the time. In practice CAT and AIC lead in many examples to exactly the same orders. It appears reassuring that quite different conceptual foundations can lead to similar conclusions in practice.

An important application of fitting an approximating autoregressive scheme $AR(\hat{p})$ to a time series is the estimation of information numbers which are used to determine the goodness of fit of ARMA (p,q) schemes. It should be emphasized that the ultimate decision on the adequacy of a model should be based on a definition of "parsimony of a model" which requires that the spectral distribution function of the residuals $(Y|\text{variables in model})^{\nu}(t)$ be "parsimoniously" not significantly different from white noise.

The spectral density of the memory p prediction errors $Y^{\nu,p}(\cdot)$ can be expressed in terms of $g_p(z)$, the autoregressive transfer function of order p, by

$$f_{Y^{\nu,p}}(\omega) = \frac{1}{\sigma_p^2} |g_p(e^{2\pi i\omega})|^2 f(\omega)$$

If the time series $Y(\cdot)$ is in fact AR(p), then its spectral density equals the <u>approximating autoregressive spectral density</u> $f_p(\omega) = \sigma_p^2 |g_p(e^{2\pi i\omega})|^{-2}$

A time series $Y(\cdot)$ can be regarded as approximated by an AR(p) if

$$\bar{f}_p(\omega) = \frac{f(\omega)}{f_p(\omega)}$$

can be regarded as not "significantly" different from a constant. In this way a test of the hypothesis that a time series $Y(\cdot)$ is AR(p) can be converted to a test of the hypothesis that the prediction error time series is white noise.

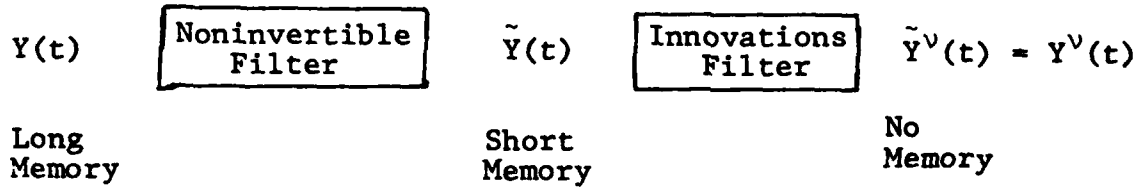
Extremely useful diagnostics concerning model identification and fit are provided by spectral distribution functions

$$F(\omega) = 2 \int_0^{\omega} f(\omega')\, d\omega', \qquad 0 \leq \omega \leq 0.5$$

6. Model Identification of Long Memory Time Series

A time series is diagnosed as being long memory when it is not no memory or short memory. There are two important models for a long memory time series $Y(t)$.

A. Transformable to a short memory time series by a non-invertible filter.



The non-invertible filter is chosen to be a difference operator of the form

$$\tilde{Y}(t) = G(L)\, Y(t) = Y(t) + A(1)Y(t-1) + \ldots + A(M)Y(t-M).$$

Then $\tilde{Y}(t)$ has sample spectral density

$$\tilde{f}_{\tilde{Y}}(\omega) = |G(e^{2\pi i\omega})|^2 \, \tilde{f}_Y(\omega)$$

Finding the non-invertible filter can be regarded as an additive decomposition of the log spectral densities: choose G so that

$$\log \tilde{f}_{\tilde{Y}}(\omega) = \log \tilde{f}_Y(\omega) + \log |G(e^{2\pi i\omega})|^2$$

looks like the sample spectral density of a short memory time series. ARARMA schemes [Parzen (1982)] choose G by best lag non-stationary autoregression.

B. Representable as the sum $Y(t) = S(t) + N(t)$ of a long memory signal plus a short memory noise.

$$Y(t) \begin{cases} S(t) \text{ long memory} \\ N(t) \text{ short memory} \end{cases}$$

A usual approach to finding S(t) is to model it as a sum of pure harmonics:

$$S(t) = \sum_{j=1}^{k} \{A_j \cos 2\pi\omega_j t + B_j \sin 2\pi\omega_j t\}$$

It is difficult to identify how many terms, and what frequencies ω_j, to include in S(t). A new approach could be based on regarding this model as an additive decomposition of the sample spectral density:

$$\tilde{f}_Y(\omega) = \tilde{f}_S(\omega) + \tilde{f}_N(\omega).$$

One defines S(t) by first forming $\tilde{f}_S(\omega)$ which is chosen so that

$$\tilde{f}_N(\omega) = \tilde{f}_Y(\omega) - \tilde{f}_S(\omega)$$

looks like the sample spectral density of a short memory time series. One determines a threshold value C which is to be subtracted from $\tilde{f}_Y(\omega)$ to form $\tilde{f}_S(\omega)$ by

$$\tilde{f}_S(\omega) = \{\tilde{f}_Y(\omega) - C\}_+ .$$

One determines the threshold value by treating $\tilde{f}(k/Q)$ as a data batch to be studied by non-parametric data modeling methods using quantile functions [see Parzen (1979)].

7. The Array of Spectral Estimators

Given a time series sample $\{Y(t), t=1,2,\ldots,T\}$ a bewildering array of estimated spectral densities $f(\omega)$ can be formed.

A. Pre-processing. To analyze a time series sample $Y(t)$, $t=1,\ldots,T$, one will proceed in stages which often involve the subtraction of or elimination of strong effects in order to see more clearly weaker patterns in the time series structure.

The aim of pre-processing is to transform $Y(\cdot)$ to a new time series $\tilde{Y}(\cdot)$ which is short memory. Some basic pre-processing operations are memory less transformation (such as square root and logarithm), detrending, "high pass" filtering, and differencing. One usually subtracts out the sample mean $\bar{Y} = \frac{1}{T}\sum_{t=1}^{T} Y(t)$; then the time series actually processed is $Y(t) - \bar{Y}$. If the mean $\bar{Y}$ is a large number, it should be subtracted; the variations in $Y(t)$ are then the variations of $Y(t)$ about its mean. The sample mean $\bar{Y}$ and sample variance $\hat{R}(0)$ should always be recorded.

B. Sample Fourier Transform by Data Windowing, Extending with Zeroes, and Fast Fourier Transform. Let $Y(t)$ denote a pre-processed time series. The first step in the analysis could be to compute successive autoregressive schemes using

operations only in the time domain. An alternative first step is the computation of the sample Fourier transform

$$\tilde{\psi}(\omega) = \sum_{t=1}^{T} Y(t) \exp(-2\pi i \omega t)$$

at an qui-spaced grid of frequencies in $0 \leq \omega \leq 1$, of the form $\omega = \frac{k}{Q}$, $k=0,\ldots,Q-1$. We call Q the spectral computation number. One should chose $Q \geq T$, and we recommend $Q \geq 2T$.

Prior to computing $\psi(\omega)$, one should extend the length of the time series by addition zeroes to it. Then $\tilde{\psi}(\omega)$, $\omega = \frac{k}{Q}$, can be computed using the Fast Fourier transform.

If the time series may be long memory one should compute in addition a sample "data windowed" Fourier transform

$$\tilde{\psi}_W(\omega) = \sum_{t=1}^{T} Y(t) W(\tfrac{t}{T}) \exp(-2\pi i \omega t) \quad .$$

To understand the effect of the window, one replaces Y(t) by a spectral representation $Y(t) = \int_0^1 \exp(2\pi i \lambda t)\, d\psi(\lambda)$; then

$$\tilde{\psi}_W(\omega) = \int_0^1 w_T(\omega-\lambda)\, d\psi(\lambda) \text{ where } w_T(\lambda) = \sum_{t=1}^{T} W(\tfrac{t}{T}) \exp(-2\pi i \lambda t).$$

Considerations involved in the choice of data windows are discussed in Harris (1978).

C. Sample Spectral Density. The sample spectral density $\tilde{f}(\omega)$ is obtained essentially by squaring and normalizing the sample Fourier transform;

$$\tilde{f}(\omega) = \frac{|\tilde{\psi}(\omega)|^2}{\frac{1}{Q}\sum_{k=0}^{Q-1} |\tilde{\psi}(\frac{k}{Q})|^2}, \quad \omega=\frac{k}{Q},\ k=0,1,\ldots,Q-1.$$

D. Nonparametric kernel spectral density estimator. An estimator $\hat{f}(\omega)$ of the spectral density is called: parametric when it corresponds to a parametric model for the time series (such as an AR or ARMA model); non-parametric otherwise. A general form of non-parametric estimator is the kernel estimator.

$$\hat{f}(\omega) = \sum_{v=-\infty}^{\infty} k(\frac{v}{M})\ \hat{\rho}(v)\ e^{-2\pi i\omega v}\ , \quad 0\leq\omega\leq 1.$$

Two popular choices of kernel are the Parzen window [Parzen (1961)]

$$\begin{aligned} k(t) &= 1-6t^2 + 6t^3 \quad , \quad |t| \leq 0.5 \quad , \\ &= 2\ (1 - |t|)^3 \quad , \quad 0.5 \leq |t| \leq 1 \quad , \\ &= 0 \quad , \quad 1 \leq |t|. \end{aligned}$$

and the spline-equivalent window [Parzen (1958), Cogburn and Davis (1974), Wahba (1980)]

$$k(t) = \frac{1}{1+t^{2r}}$$

where $r \geq 2$ is usually chosen to equal 2 or 4. The problem of determining optimum truncations points M has no general solution;

one approach is to choose a large value of M to obtain a preliminary smoothing of the sample spectral density.

E. Autoregressive spectral density estimators. The Yule-Walker equations are solved to estimate innovation variances $\hat{\sigma}_m^2$, to which are applied order determining criteria (AIC, CAT) to determine optimal orders m and also to test for white noise. The value of $\hat{\sigma}_m^2$ and the dynamic range of the autoregressive spectral estimator $\hat{f}_m(\omega)$ are used to determine the memory type of the time series [Parzen (1982)]. One should determine a best and second best AR order.

F. ARMA spectral density estimations. When a time series is classified as short memory the approximating AR scheme is used to form the MA(∞) coefficients which are used to form a subset regression procedure for determining the best fitting ARMA scheme, and the corresponding ARMA spectral density estimator.

We do not believe that spectral estimation is a non-parametric procedure to be conducted independently of model identification. The final form of spectral estimator should be based on an identification of the type (AR, MA, or ARMA) of the whitening filter of a short memory time series.

G. Non-stationary autoregression. When a time series is classified as long memory, more accurate estimators of autoregressive coefficients are provided by minimizing a "forward and backward" least squares criterion

$$\sum_{t=m+1}^{T} (Y(t) + \alpha_m(1)\ Y(t-1) + \ldots \alpha_m(m)\ Y(t-m))^2$$

$$+ (Y(t-m) + \alpha_m(1)\ Y(t-m+1) + \ldots + \alpha_m(m)\ Y(t))^2,$$

or by Burg estimators [for references to descriptions of Burg's algorithm, see Kay and Marple (1981)].

When several harmonics are present in the data, whose frequencies are close together, least squares autoregressive coefficient estimators are more effective than Yule-Walker autoregressive coefficient estimators in providing autoregressive spectral estimators which exhibit the split peaks one would like to see in the estimated spectral density.

H. Spectral density estimators based on inverse correlations and cepstral correlations. Additional insight into the peaks and troughs to be given significance in the final estimator of the spectrum of a short memory time series can be provided by forming nonparametric kernel estimators of $f^{-1}(\omega)$ and $\log f(\omega)$.

For a spectral density $f(\omega)$ obeying suitable conditions, one can define the inverse-correlation function [see Cleveland (1972), Parzen (1974), Chatfield (1979)]

$$\rho i(v) = \int_0^1 e^{2\pi i v\omega}\ f^{-1}(\omega)\ d\omega \div \int_0^1 f^{-1}(\omega)\ d\omega$$

and the cepstral-correlation function [see Wahba (1980) for an application]

$$\gamma(v) = \int_0^1 e^{2\pi i v\omega} \log f(\omega)\, d\omega$$

It should be noted that the inverse-correlation function is non-negative definite. However the cepstral-correlation function is not. These new types of correlation functions are introduced because they may provide more parsimonious parametrizations in the sense that they decay to 0 faster than does the correlation function. Statistical inference (from a sample) of the probability law of a time series often achieves greatest statistical efficiency by using the most parsimonious parametrizations. Thus to form estimators $\hat{f}(\omega)$ of the spectral density $f(\omega)$ from a raw estimator $\tilde{f}(\omega)$, greater precision may be attained by first forming estimators $\{f^{-1}(\omega)\}\hat{}$ and $\{\log f(\omega)\}\cdot$ of the inverse or logarithm of the spectral density. Autoregressive spectral estimation may be regarded as an approach to estimating $f(\omega)$ by first estimating $f^{-1}(\omega)$ [Durrani & Arslanian (1982)].

8. Quantile approach to non-Gaussian data analysis

A promising new approach to the analysis of non-Gaussian time series analysis provided by the probability density-quantile function introduced by Parzen (1979). This section should be of theoretical interest to spectral analysis because of the many isomorphisms that exist between the statistical theory of spectral density estimators and probability density-quantile estimators. It should be of interest as a technique to answer statistical questions about non-Gaussian time series for which correlation functions are not an appropriate tool. Here we have only space to mention some basic definitions.

The probability distribution of a continuous random variable Y is most insightfully described not by its distribution function $F_Y(y) = Pr[Y \leq y]$ and its probability density $f(y)=F'(y)$, but by its quantile function

$$Q_Y(u) = F_Y^{-1}(u) = \inf\{y:F_Y(y) \geq u\}$$

and its density quantile function

$$f_Y Q_Y(u) = f_Y(Q_Y(u)).$$

The bivariate distribution $f_{Y(t),Y(t+v)}(y_1,y_2)$ is best described by the joint density-quantile function

$$f_{Y(t),Y(t+v)}\ (Q_{Y(t)}(u_1),\ Q_{Y(t+v)}(u_2)).$$

Techniques for estimating a joint density-quantile function are currently under development by Woodfield, using the dependence density $d(u_1,u_2)$ defined for two random variables X and Y by

$$f_{X,Y}(Q_X(u_1),\ Q_Y(u_2)) = d_{X,Y}(u_1,u_2)f_X Q_X(u_1) f_Y Q_Y(u_2)$$

Estimators of $d_{X,Y}(u_1,u_2)$ are obtained from raw estimators of

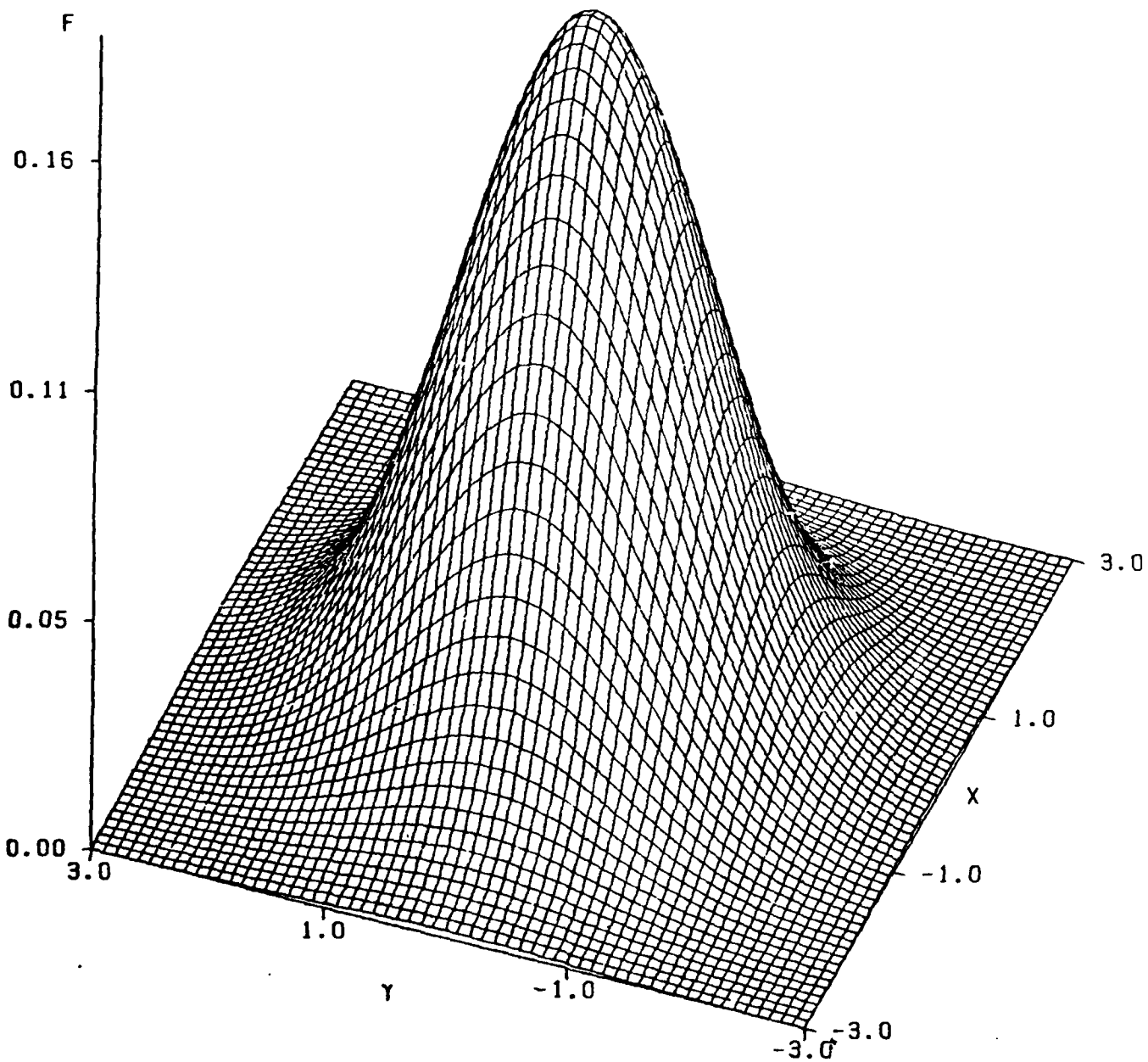
$$D_{X,Y}(u_1,u_2) = F_{X,Y}(Q_X(u_1),Q_Y(u_2))$$

which is the joint distribution function of

$$U_1 = F_X(X), \quad U_2 = F_Y(Y)$$

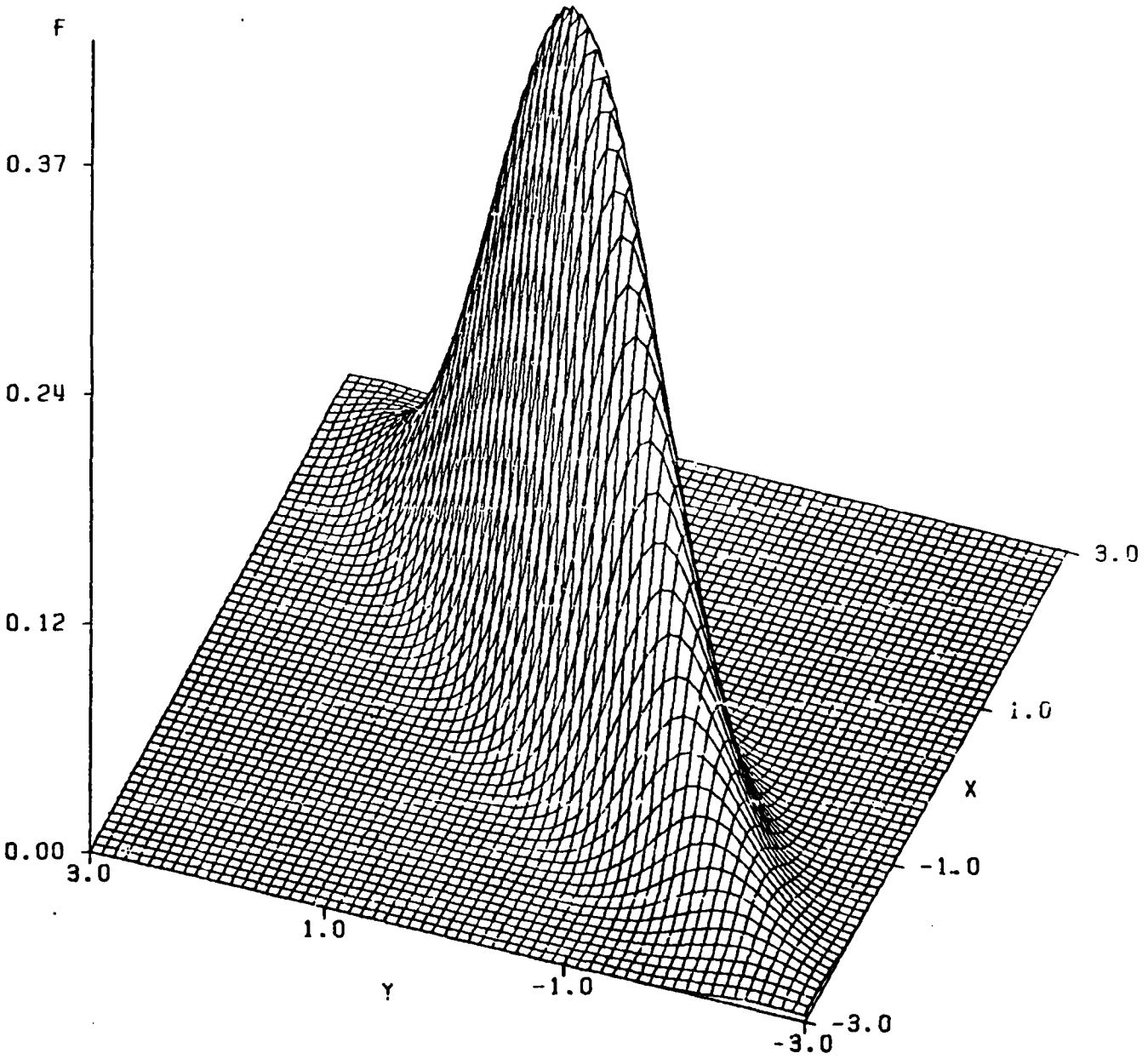
Information criteria can be used to test the independence of X and Y, and to estimate their joint density. Note that $I(Y|X) = I(f_{X,Y};f_X f_Y) = -H(d_{X,Y}(u_1,u_2))$

BIVARIATE DATA ANALYSIS

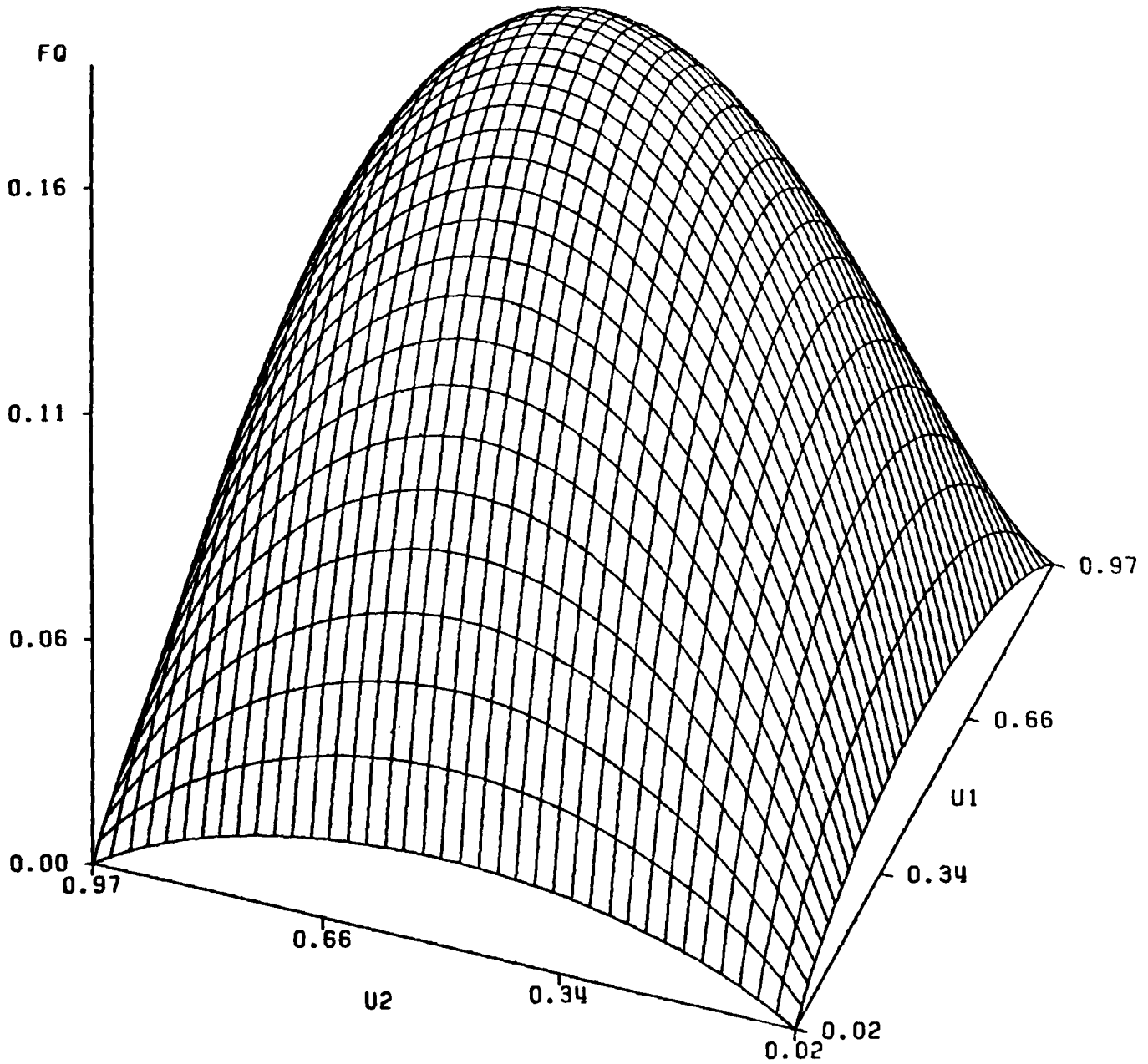


F(X,Y) FOR BIVARIATE NORMAL CASE WITH RHO=0.0

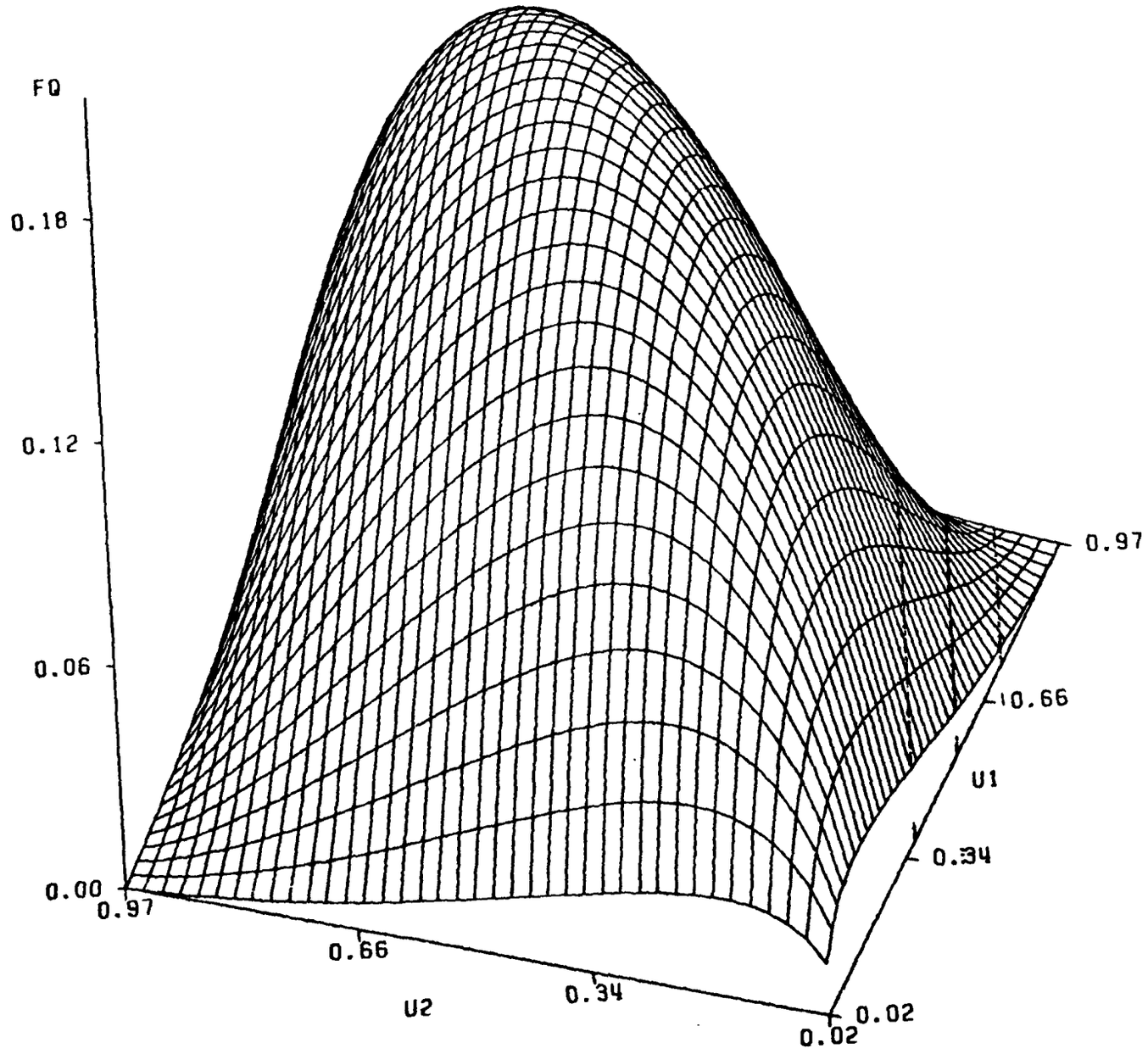
BIVARIATE DATA ANALYSIS



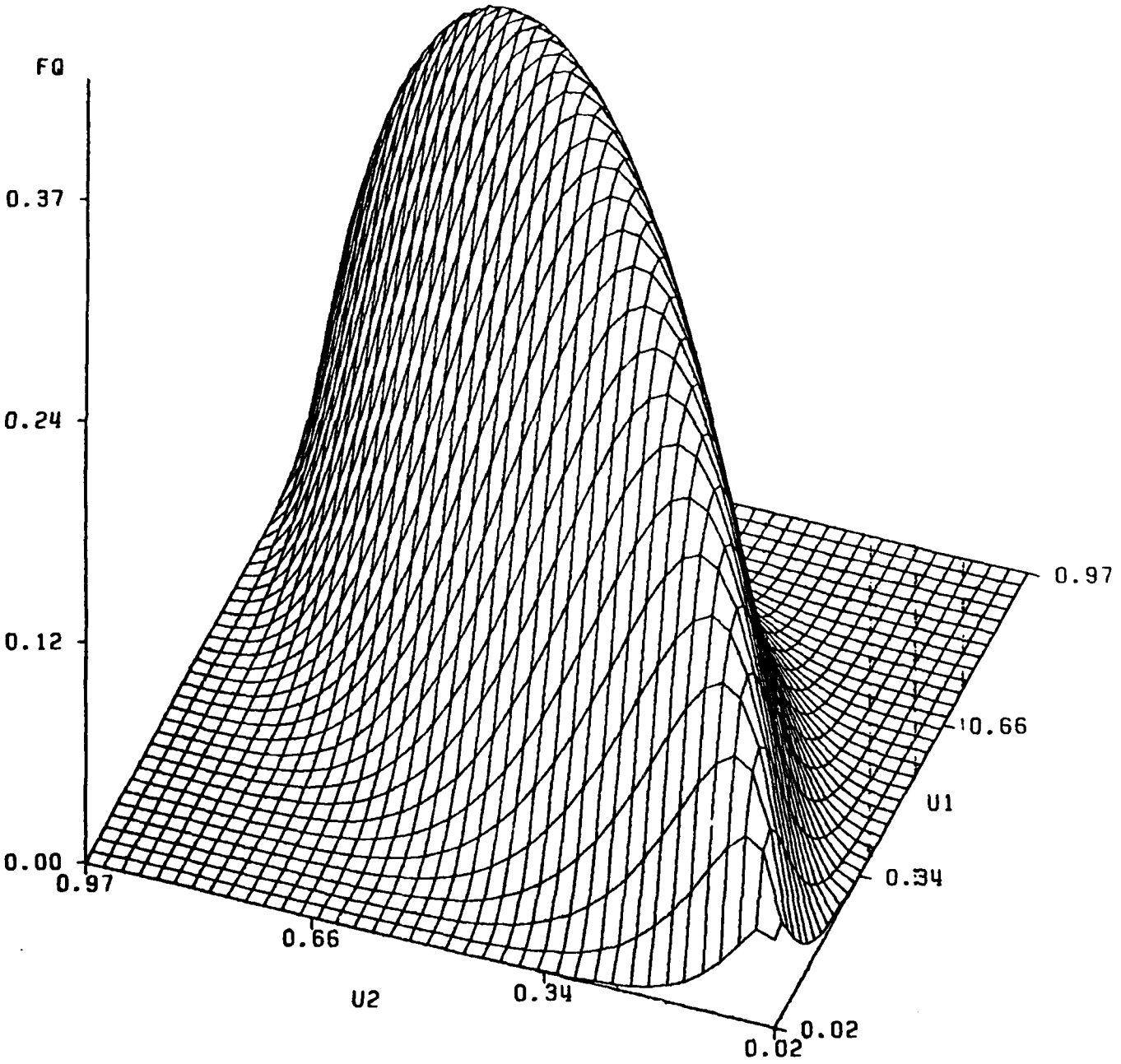
F(X,Y) FOR BIVARIATE NORMAL CASE WITH RHO=0.9



DENSITY-QUANTILE FOR BIVARIATE NORMAL CASE WITH RHO=0.0



DENSITY-QUANTILE FOR BIVARIATE NORMAL CASE WITH RHO=0.5



DENSITY-QUANTILE FOR BIVARIATE NORMAL CASE WITH RHD=0:9

REFERENCES

Akaike, H. (1970) A fundamental relation between predictor identification and power spectrum estimation. Ann. Inst. Statist. Math. 22, 219-223.

_____. (1974) A new look at the statistical model identification, IEEE Trans. Autom. Contr., AC-19, 716-723.

_____. (1977) On entropy maximization principle, Applications of Statistics, P. R. Krishnaiah, ed., North-Holland, Amsterdam, 27-41.

Beamish, M. and Priestley, M. B. (1981) A study of autoregressive and window spectral estimation. Appl. Statist., 30, 41-58.

Berk, K. N. (1974) Consistent autoregressive spectral estimates. Ann. Statist., 2, 489-502.

Blackman, R. B. and Tukey, J. W. (1959) The Measurement of Power Spectrum from the point of view of Communication Engineering. Dover: New York. (Originally published in Bell Systems Tech. Journal, 37, 185-282 and 485-569 (1958)).

Burg, J. P. (1968) A new analysis technique for time series data, NATO Advanced Study Institute on Signal Processing, Enschede, The Netherlands.

Carmichael, J. P. (1976) The Autoregressive Method, Ph.D. thesis, Statistical Science Division, State University of New York at Buffalo.

_____. (1978) Consistency of the Autoregressive Method of Density Estimation, Technical Report, Statistical Science Division, State University of New York at Buffalo.

Carter, G. C. and Nuttall, A. H. (1982) An analysis of the generalized spectral estimation technique, including a near-optimum method for sonar applications. Proceedings IEEE. To be published.

Chatfield, C. (1979) Inverse autocorrelations, J.R. Statistic. Soc.A 142,363-77.

Childers, D. G. (1978) Modern Spectrum Analysis, New York: IEEE Press.

Cleveland, W. S. (1972) The inverse autocorrelations of a time series and their applications. Technometrics 14, 277-93.

Cogburn, R. and Davis, H. T. (1974) Periodic splines and spectral estimation. Annals of Statistics, 2, 1108-1126.

Durrani, T. S. and Arslanian, A. S. (1982) Windows associated with high resolution spectral estimators. Submitted for publication.

Goodman, H. R. (1957) On the joint estimation of the spectra, cospectrum, and quadrative spectrum of a two-dimension stationary Gaussian process, Ph.D. Thesis, Department of Statistics, Princeton University.

Grenander, U. (1981) Abstract Inference New York: Wiley.

Hannan, E. J. and Quinn, B. G. (1979) The determination of the order of an autoregression, Journal of the Royal Statistical Society, 41, 190-195.

Harris, F. (1978) On the use of windows for harmonic analysis with the discrete Fourier transform, Proc. IEEE, 66, 51-83.

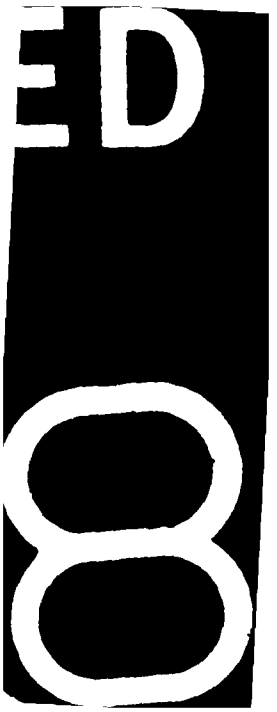
Kay, S. M. and Marple, Stanley L., Jr. (1981) Spectrum Analysis--A Modern Perspective. Proceedings of the IEEE, 69, 1380-1419.

Kromer, R. E. (1969) Asymptotic properties of the autoregressive spectral estimator. Ph.D. Thesis. Statistics Dept., Stanford University.

Munk, W. H. and Snodgress, F. E. (1957) "Measurements of southern swell at Guadalupe Island," Deep-Sea Research, 4, 272-286.

Nevai, Paul G. (1979). An asymptotic formula for the derivatives of orthogonal

polynomials. SIAM J. Math. Anal. 10, 472-477.
Pagano, M. (1980) Some recent advances in autoregressive processes. Directions in Time Series, ed. D. R. Brillinger and G. C. Tiao, Institute of Mathematical Statistics, 280-302 (Comments by H. T. Davis).
Parzen, E. (1957) "On consistent estimates of the spectrum of a stationary time series" Ann. Math. Statist., 28, 329-348.
_____. (1958) "On asymptotically efficient consistent estimates of the spectral density function of a stationary time series" J. Roy. Statist. Soc., B., 20, 303-322.
_____. (1961) "Mathematical considerations in the estimation of spectra" Technometrics, 3, 167-190.
_____. (1964) "An approach to empirical time series analysis" J. Res. Nat. Bur. Standards, Sect. D, 68D, 937-951.
_____. (1967) "Empirical multiple time series analysis" Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, ed. L. LeCam, and J. Neyman, University of California Press, Vol. I, pp. 305-340.
_____. (1967) "The role of spectral analysis in time series analysis" Review of the International Statistical Institute, 35, 125-141.
_____. (1968) "Statistical spectral analysis (single channel case) in 1968" Proceedings of NATO Advanced Study Institute on Signal Processing, Enschede, Netherlands.
_____. (1969) "Multiple time series modeling" Multivariate Analysis-II, ed. P. Krishnaiah, Academic Press: New York, 389-409.
_____. (1970) "Statistical inference on time series by RKHS methods, II" Proceedings 12th Seminar Canadian Mathematical Congress, ed. R. Pyke, Canadian Mathematical Congress, Montreal 1-37.
_____. (1974) "Some Recent Advances in Time Series Modeling" IEEE Transactions on Automatic Control, AC-19, 723-730.
_____. (1977) "Multiple Time Series: Determining the Order of Approximating Autoregressive Schemes" Multivariate Analysis-IV, Ed. P. Krishnaiah, North Holland: Amsterdam, 283-295.
_____. (1979) "Nonparametric Statistical Data Modeling" Journal of the American Statistical Association, (with discussion), 74, 105-131.
_____. (1979) "A Density-Quantile Function Perspective on Robust Estimation" Robustness in Statistics, ed. R. Launer and G. Wilkinson, New York: Academic Press, 237-258.
_____. (1980) "Time Series Modeling, Spectral Analysis, and Forecasting" Directions in Time Series Analysis, ed. D. R. Brillinger and G. C. Tiao, Institute of Mathematical Statistics.
_____. (1981) Autoregressive Spectral Estimation, Log Spectral Smoothing, and Entropy. IEEE Spectral Estimation Workshop (Hamilton, Canada).
_____. (1982) "Time Series Model Identification and Prediction Variance Horizon," Applied Time Series Analysis II, ed. D. Findley, New York: Academic Press, 415-447.
_____. (1983) Autoregressive Spectral Estimation, Handbook of Statistics III, ed. D. Brillinger and P. Krishnaiah, Amsterdam: North Holland.
VanDenBos, A. (1971) Alternative interpretation of maximum entropy spectral analysis, IEEE Trans. Inform. Theory, IT-17, 493-494.
Wahba, Grace (1980). Automatic smoothing of the log periodogram, Journal of the American Statistical Assn., 75, 122-132.
Wiener, N. (1930) Generalized harmonic analysis. Acta. Math., 55, 117-258.
Parzen, E. (1982) ARARMA Models for Time Series Analysis and Forecasting, Journal of Forecasting 1, 67-82.
Pinsker, M. (1963) Information and Information Stability of Random Variables, Holden Day: San Francisco.


ED
8